



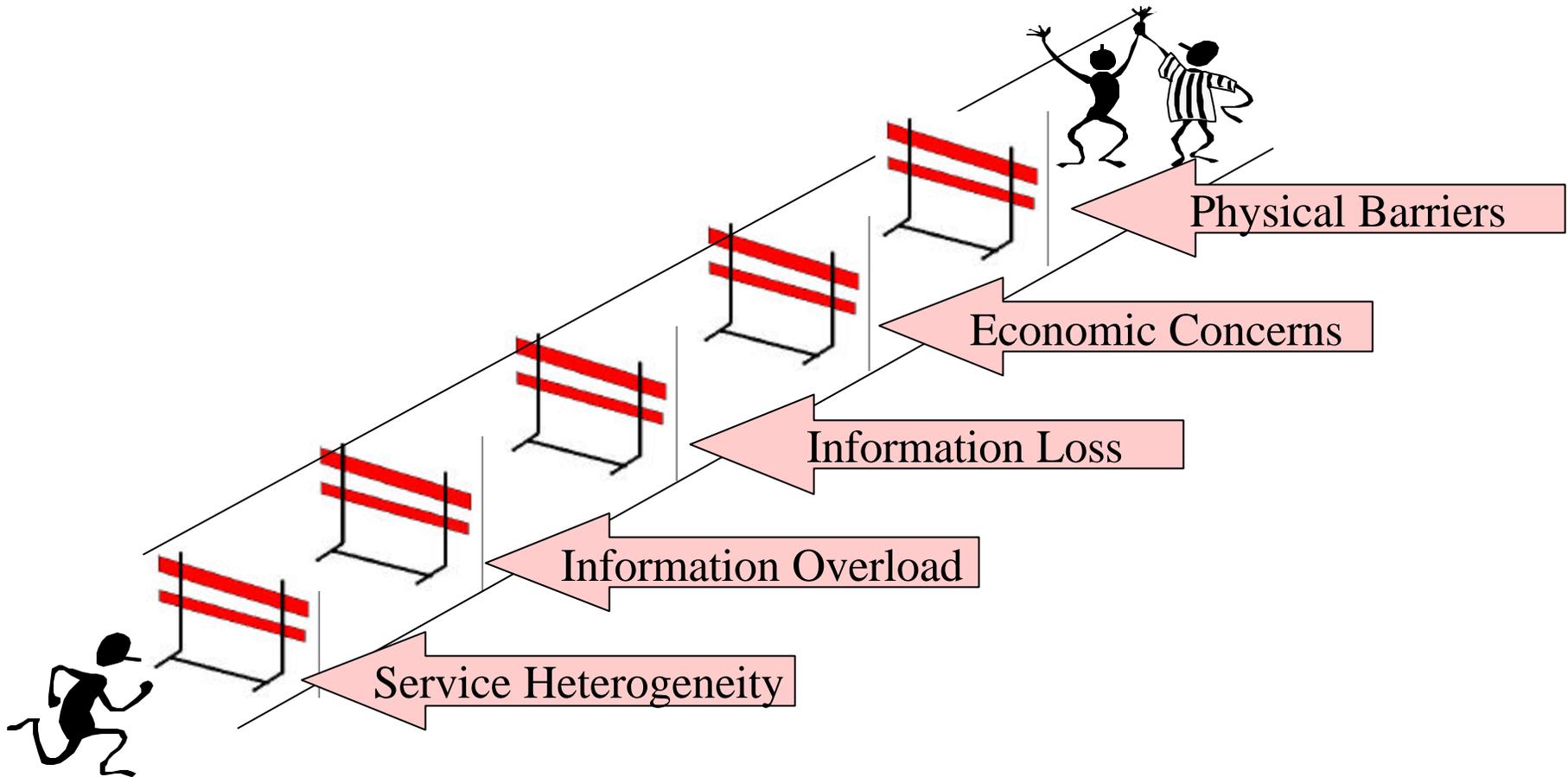
Stanford InterLib Technologies

Hector Garcia-Molina
and the Stanford DigLib Team

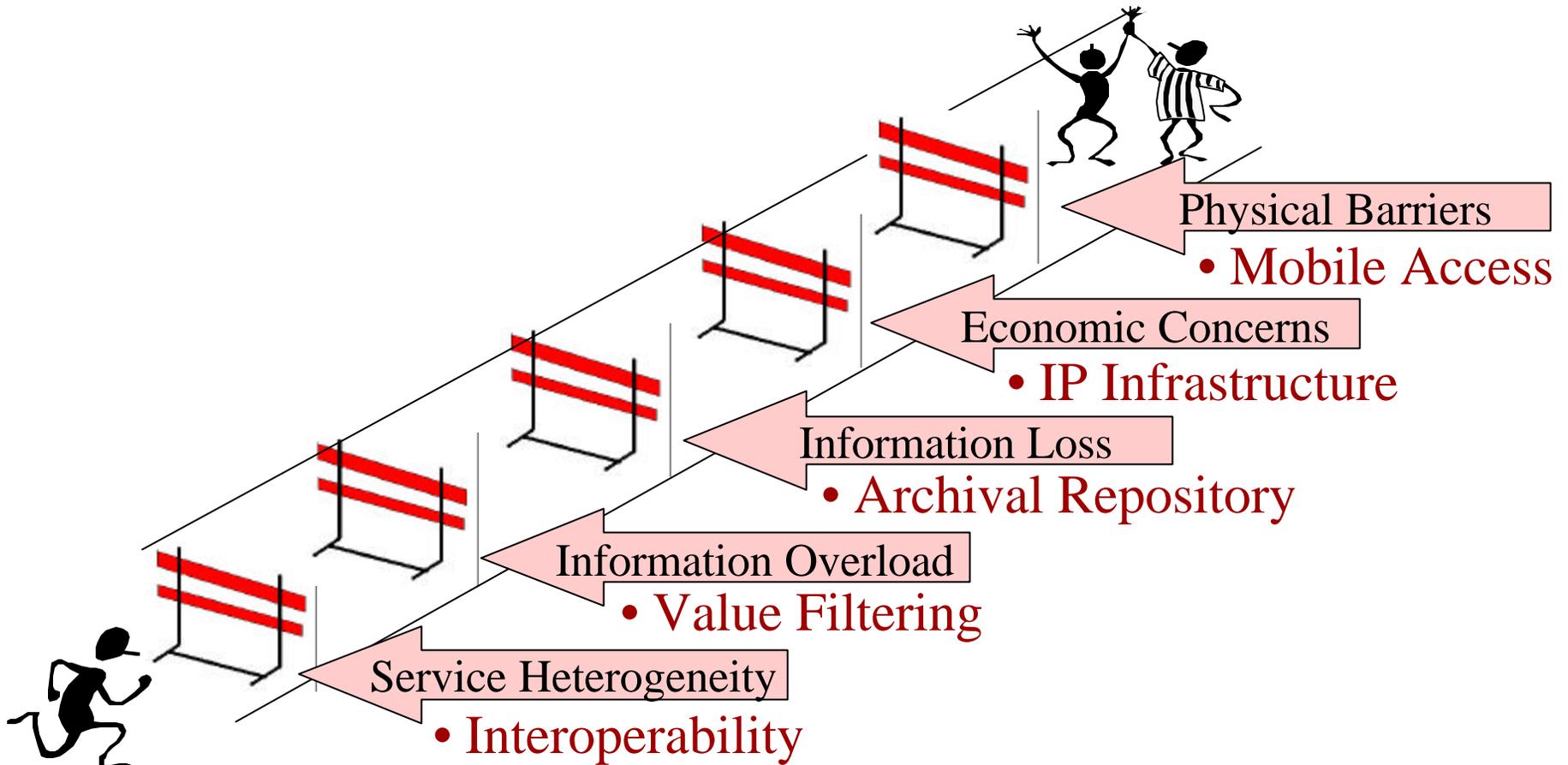
Stanford Digital Libraries Team

- Faculty:
 - Dan Boneh, Hector Garcia-Molina, Terry Winograd
- Research Scientist
 - Andreas Paepcke
- Librarians
 - Vicky Reich, Rebecca Wesley
- Partners:
 - InterLib Partners, ACM, Dialog, Hitachi, IBM, Intel, Microsoft, NASA Ames Library, Stanford Libraries, SUL HighWire Press, Xerox

Barriers to Effective DLs



Thrusts



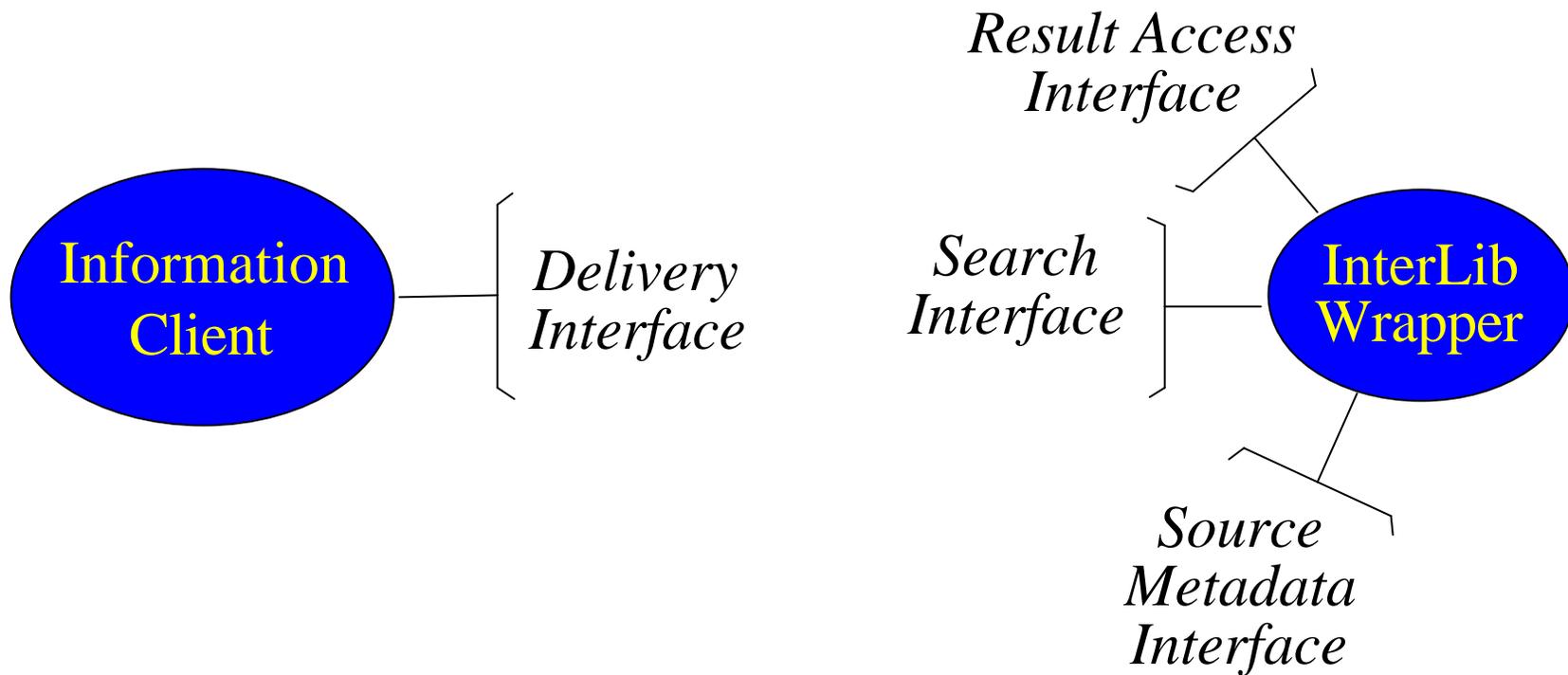
SDLIP

- Simple Digital Library Interoperability Protocol
- Goal: get InterLib (and DLI2) to interoperate!!

Search Protocol: Initial Goals

- Trivial to implement!
 - Works over CORBA/COM, DASL/HTTP
 - Use XML
 - Does not prescribe query format
 - Does not prescribe result format
- But lets you say what you're using*
- Small footprint (Desktop/Laptop/PDA)
 - Allows for stateful or stateless operation

Interface Consists of Four Components



SDLIP Status

- Design Meeting June 22, 1999



SDLIP Status

- Design Meeting June 22, 1999



- Client & Server Toolkits Available
- Extensive Documentation
- See

<http://www-diglib.Stanford.EDU/~testbed/doc2/SDLIP/>

Current SDLIP Sources

- Some Web sources
 - People Lookup: www.switchboard.com
 - Altavista
 - IMDB (movies)
- NCSTRL services: www.ncstrl.org
 - Dienst compliant services, e.g., CoRR?
- Z39.50 servers
 - e.g., Library of Congress
- Stanford WebBase
- CDL
 - e.g., MELVYL gateway
- DASL-compliant servers

Existing Clients

- Java
 - command line
 - applet
- C++
 - Palm Pilot
- TCL (Ray Larson)
- DASL-compliant clients

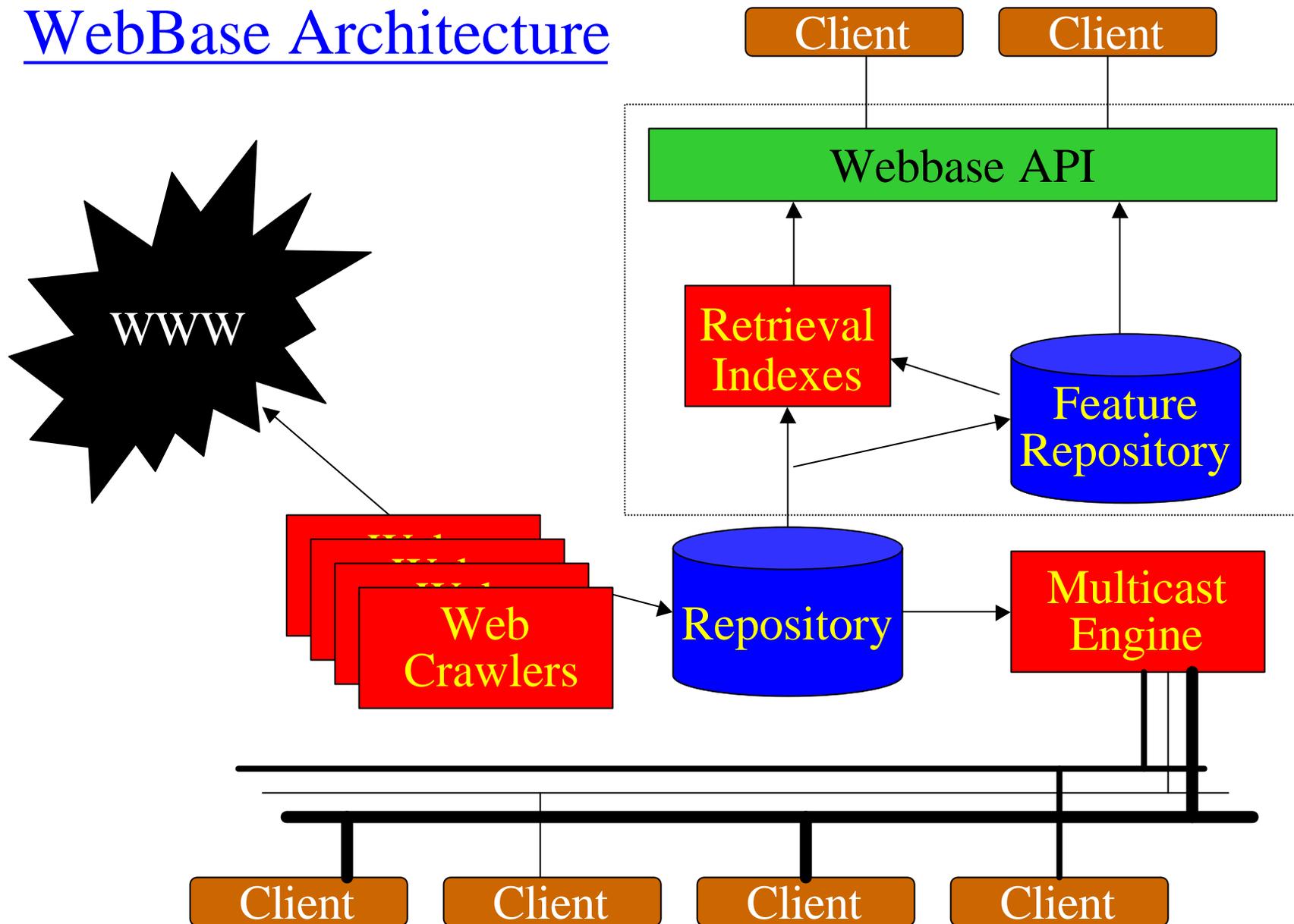
WebBase Goals

- Manage very large collections of Web pages
- Enable large-scale Web-related research
- Locally provide a significant portion of the Web
- Efficient wide-area Web data distribution

Challenges

- Huge information space
 - Wide area distribution
 - URL space (to remember while crawling)
 - Web content (to store)
- Limited resources
 - Disk
 - Time
 - Memory
 - Bandwidth
 - Server administrator tolerance
- Continuous evolution
 - More pages
 - Pages change/disappear
 - Mirror sites installed
 - Keeping data “fresh”
- Crawling issues
 - Data ‘fiefdoms’: firewalls; access permissions; load controls
 - Overhead per site: DNS lookups; processing robots.txt
 - Parallelization
 - Ability to interrupt & restart

WebBase Architecture



Stanford InterLib Technologies

